DOCUMENT RESUME

ED 274 695 TM 860 563

AUTHOR Goldstein, Harvey; Wolf, Alison
TITLE Practical Testing on Trial: A Study of the Reliability and Comparability of Results under Decentralized System of Practical Assessment.
PUB DATE Apr 86
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; *Criterion Referenced Tests; *Evaluation Criteria; Foreign Countries; High Schools; Interrater Reliability; *Performance Tests; Teacher Made Tests; Test Construction; Test Format; Testing Problems; *Test Reliability; Vocational Education; *Work Sample Tests
IDENTIFIERS United Kingdom

ABSTRACT

Locally developed occupational tests were administered to 16- and 17-year-olds in a government-sponsored vocational education program in the United Kingdom over a six-month period in 1984. Job skills were tested in two occupational areas: use of a micrometer and invoice completion. Some performance tests were designed by researchers and some by the trainees' supervisors, who administered the tests. Standardized national tests were not used. One purpose of this study was to examine the feasibility and implications of this system of decentralized testing, in which practical tests were designed and administered following central guidelines. The other major purpose of this study was to examine the factors affecting test reliability, within-site examiner reliability, and the comparability of test content and standards across sites. Results indicated that of the examiners who were inconsistent in their application of judgment criteria, five percent were inconsistent in relation to their own micrometer exercises, and 45 percent were inconsistent with their own invoice tests. There appeared to be considerable variation in the standards used between different raters. Results from studying the internal reliability estimates revealed the dangers of a single assessment. The trainees' behavior was extremely variable when performance on two different tasks was assessed. Appendixes describe the methodology more fully and provide formulae for reliability coefficients. Thirteen tables are offered. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED274695

Practical Testing on Trial: A Study of the Reliability and Comparability of Results under Decentralised System of Practical Assessment

HARVEY GOLDSTEIN
ALISON WOLF

University of London Institute of Education

ABSTRACT

The paper reports on the results of a study of decentralised practical assessment, carried out with 16 and 17 year olds enrolled in government-sponsored vocational education programmes. The study was designed to:

(1) provide general information on the feasibility and implications of assessing students using a decentralised system in which practical tests were designed and administered locally following central guidelines;

(2) estimate and examine the factors affecting test reliability; within-site tester reliability; and the comparability of test content and standards across site;

(3) investigate the policy implications of (1) and (2) above, especially for the type of oversight and moderating system that should be established by the central bodies responsible for assessment.

Most test procedures used for selection or certification involve the use of "alternate forms", within a given time period and/or across years. However, very little information is available on the reliability and comparability of such supposedly parallel instruments. The study was funded in order to provide such information, and specifically to do so using practical tests in which occupational validity and not discrimination between subjects, determined which items were included.

TM 860 563

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. Wolf

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Practical Testing on Trial: A Study of the Reliability and Comparability of Results under a Decentralised System of Practical Assessment

Paper presented to the Symposium on "Research on Mandated Testing": American Educational Research Association Annual Meeting, 1986

Alison Wolf
University of London Institute of Education

Introduction

Education in the United Kingdom currently is undergoing one of its periodic and major reorganisations. To a large degree, this reorganisation is assessment-led: that is, various forms of examination or assessment are being used quite deliberately to affect the nature of secondary schooling, and of vocational training, and also the relationship between children's education and their later careers. Whether the changes that result will be those envisaged by their architects is doubtful. That there will be major changes is not.

If one outlines the two most evident areas of activity, interesting parallels emerge with current developments in the United States. The first is the almost universal espousal of "criterion referencing" and the attempt to make the whole system of public examinations criterion referenced. In the UK and other European countries, school level assessments are set and marked by central, governmental or quasi-governmental bodies; and success determines progress into further education and into the job market. A similar, though rather less tightly structured system obtains for vocational and technical qualifications, many of which are taken by young people who leave school at 16, and study part time while in employment or on government-sponsored training schemes. Scottish examinations - both school exams for 16 year olds, and the whole vocational system - are now supposedly fully criterion referenced. The new English General Certificate of Secondary Education is moving towards what are labelled "grade criteria". Finally, in vocational training, where there has been a huge increase in government spending, the agency concerned, the Manpower Services Commission, is alternatively coaxing and whipping the accrediting bodies and the employers' associations into creating a comprehensive system of "competency based" assessment.

The second major development is the decentralisation of assessment. Public exams - including many vocational and technical assessments - have tended to be pen and paper, with candidates sitting identical papers under carefully controlled conditions, and with marking done centrally. Even with the broader "essay type" questions which are so important throughout Europe, this limits severely what can be assessed. The validity of such exams is under increasting attack. Consequently, there is growing interest in and movement towards a system in which local curricula, continuous assessment, and practical

projects play a part. Once again, this development holds for both school examinations and vocational assessment.

A British observer of American education is immediately struck by your similar growth industry of criterion referenced competency testing. In both countries there seems to be a strong popular sentiment that we can and should be able to define what someone in a particular job, or "owning" a particular skill, should be able to do. Then we should test them to see if they can in fact do it. However, most of your educational competency tests - including those for teachers - appear to be rather traditional in form, relying heavily on pen and paper exercises, centrally set and marked. The parallels with our other development, of increasingly decentralised, practical, teacher or trainer-led assessments, are found more clearly in your training programmes.

The Job Training Partnership Act's call for "performance standards", and for "employment competencies" recognised by "private industry councils", sounds very familiar to any observer of our Manpower Services Commission. Here, as at home, we find that the aims are decentralised assessments of defined goals; the accreditation of practical skills; and the opening up of educational and job opportunities which in the past depended on traditional tests and grades.

The shifts I have described are in large part a response to worries about validity; concern about a lack of skills among many school leavers; and a desire to open up opportunities to the many who fail in our traditional academic curriculum. However, they are accompanied by concerns about comparability, maintaining standards, and how best to combine flexibility with fairness. The University of London Institute of Education was therefore commissioned to study how assessments of applied skills would actually work in a decentralised, teacher or trainer-led system. The research, described here, was conducted for the Manpower Services Commission, whose remit is vocational training; and involved young people in training schemes, not full time education. The results concern test content, assessment standards, and the consistency of student or trainee performance on competency tests. The first two topics are of most direct relevance in contexts, such as JTPA programmes, where centralised testing is not involved; although they also raise interesting questions about the consistency of course content and grading by teachers generally. Evidence on the latter topic, however, is of great relevance to any situation involving competency testing - and the more so, the more "high stakes" the test.

Research Design *1*

477 trainees and 98 supervisors were directly involved in the research, which was conducted during 1984-85. Essentially the work comprised two separate but related studies. One looked at the results of giving trainees two assessments which were extremely similar, but contained different numbers. A group of such exercises was written by the researchers, and represent "alternate forms" of the same exercise. This part of the research was intended to show how far apparently small differences between tests might affect assessment results.

The second study approximated far more closely to the conditions one might expect when mandated testing is carried out using locally designed or chosen assessments. Youth Training Schemes, carrying out government funded vocational training, were asked to assess trainees using both their own assessment exercise, and one of the researchers. The latter were drawn from the bank of exercises written for the alternate forms study, in this instance served as moderating exercises. The skills to be tested were specified for all participating schemes; but the sample was also subdivided, with one half being given very precise specifications, including success criteria, and the other simply a description of the skill to be tested. The two exercises which each trainee was given can again be seen as "alternate forms", but they were far less obviously related, and we refer to the results as the "own exercise" study data.

Both parts of the research were conducted in two very different occupational areas: micrometer use and invoice completion. They were chosen because we wanted to include two large but very different occupational areas, in which we could sample over a wide geographical area. The skills themselves were also very different. The former was selected as a representative of measurement tasksand as a task requiring manipulation or manual dexterity, and the latter as a characteristic task in a very wide range of clerical and retail jobs. Both also allowed for ease in scoring. Thus, a supervisor can give a definitive width for a piece of metal to be measured; and an invoice involves calculations with unique correct answers. While many tasks are not of this type, we felt that for research purposes they were preferable because coding could be relatively unproblematic and objective.

The trainees who were assessed were all involved in work or training programmes where they carried out the task concerned. All assessments were conducted by trainees' supervisors. We used a very few college lecturers in occupational areas where training had always had a strong college based component. The vast majority of supervisors, however, were not college based, although a good number did have some educational or training experience in the past. All of them were asked for a preassessment of each trainee's competence, based on whatever formal or informal method of continuous assessment the supervisors had been using to date. They were then asked for a separate assessment after each of the two exercises. In addition to using these judgements, the researchers also marked the exercises

themselves, using simply the number of items correct.2

TEST CONTENT

Any system of decentralised assessment tends to rely heavily on written instructions. This is certainly true in Britain, where both reform of school examinations and the development of a vocational system embracing all 16 year old school leavers, are generating mountains of paperwork. However we found that, almost without exception, supervisors did not read all or even any of the instructions provided. Such failure is no doubt associated with the fact that we were dealing at all times with substantive areas which they knew, and in which, indeed, they could be considered experts: but this, after all, is as it should be. Certainly our experience indicates that very little can be done to mould or change assessment procedures using written instructions. Other methods (eg videos) might be more effective, but would also be more expensive and more difficult to organise.

The problems created by people's failure to read instructions are most evident in our experience with those schemes which were asked to construct their own exercises. 3 As explained above, half these schemes were asked simply to construct an exercise which would determine whether the trainee could complete a given task satisfactorily. The other half were given considerably more detailed task descriptors, which included specified success criteria. (The exact instructions given to schemes are duplicated in Appendix I.) It was clear that a considerable number of the group which was given the more detailed descriptor either ignored or simply did not read it.

Such failure is, obviously, only apparent for items or procedures which supervisors did not include of their own volition. In the case of micrometer usage, this applied most notably to the researchers' request that the exercise should test "understanding and application of tolerances". Only about a third of those who were asked to include this, did so.

In the case of invoices, it was the supervisors' success criteria which diverged more markedly from the descriptor provided. We asked that a 100% accuracy criterion should be used - ie that the invoices should be filled in correctly, and the calculations all be correct. Such a stringent criterion was used because it is what industry representatives have described as the workplace standard in their discussions with government certifying agencies. It was, therefore, being used by the government in its piloting of standardised assessment tasks. However, in our research, well under a third of those who were given task instructions specifying this standard actually used it; while for invoice supervisors as a whole, the proportion was lower still.4 This situation is not, it should be emphasised, peculiar to workplace supervisors. Teachers are also inclined not to read instructions and the more so the more familiar the field in which they are working. 5

Failure to read instructions is less important the more supervisors in

a given industry or area have a shared understanding of what a given competence involves, and how it may be observed or assessed. Although we only examined two tasks or competencies, the results suggest strongly that there will be enormous variation in how far this is the case. Table 1 shows the content of schemes' own exercises.

TABLE 1 ABOUT HERE

It is apparent that there is greater variation within the invoicing group, both in the type of task and the number of items included in an exercise. In the case of micrometer use, it seems reasonable to talk of a shared task definition in the sense that all exercises included measurement, whether or not they included other items. In the case of invoicing, it does not. Equally, there seems no reason to suppose, on theoretical grounds, that one or other task is more 'typical'. The degree to which common task definitions exist is, rather, an empirical question, and the situation is also likely to be one of constant flux.

It is possible that an assessment system which defines vocational tasks clearly may, over time, create common task definitions where none existed before. However, this too is an empirical rather than a theoretical question; and the number of people involved and the reluctance to read instructions imply a slow process at best. Extrapolating to the situation in the schools, we find that, In England, centrally set examinations have in the past been the main force creating consensus on what a subject "is". At secondary school and college level, most teachers feel that sample exam papers, rather than the syllabus, provide the crucial information about what to teach. A shift away from this, to locally set assessments based on syllabi alone, is likely to increase the influence of the textbooks on curriculum content. Our research, it should be remembered, concerned skills as apparently specific and unambiguous as 'micrometer use' or 'invoice completion'. What can we expect from assessments of ability to 'draw an apt inference from a key statement', or 'marshall ideas and evidence in support of an argument' when speaking. 6

THE CONSISTENCY OF SUPERVISORS' JUDGEMENTS

Questions of fairness occur in similar form in any assessment system - norm or criterion referenced, external or internal, continuous or summative - and relate both to 'intra-assessor' variation and to variation between assessors and between assessment instruments. In the former case, we are concerned with whether an assessor applies the same criterion to all the individuals he or she assesses. In the latter, we are concerned with the comparability of standards used in e.g. different locations or at different times. In the former case, one's desired level of consistency will always be very high. In the latter, it may vary with circumstance. However, both the British government and the American are preoccupied with national standards, implying that standards, as well as, presumably, the content of assessments should vary extremely little over space or time.

In our research, we examined how consistently individual supervisors

interpreted assessment results in passing judgement on trainees. We were able to collect evidence relating to the success criterion each used, and the degree to which they judged a trainee's performance independently of any prior conceptions they might have held. Both these factors would affect how far trainees in the same scheme were treated alike. We also looked at 'inter-assessor variation' - i.e. the extent to which trainees were being assessed in the same way by different schemes, or the comparability of their assessment procedures.

(1) Do individual supervisors use a consistent standard in judging competence?

The judgements with which supervisors provided us were coded on a 4 point scale: 'can do' (=1); 'Maybe can' (=2); 'Maybe cannot' (=3); and 'Cannot' (=4). We compared each supervisor's rankings of their trainees with the 'raw scores' trainees obtained on an exercise - that is, how many items on an exercise were correct. Where appropriate, we also compared them with what we term 'alternative scores'. In a multi-stage operation - such as filling in an invoice - an error early in the operation will often mean that many later steps or calculations are also wrong, even if no further errors are made. Supervisors would sometimes comment to us, when grading exercises, that "There was just that one mistake throwing them off"; and 'alternative scores' only penalised an original error, not incorrect items which followed from it.

Finally, if a supervisor's judgements did not appear to embody any consistent cut offs in terms of either raw or alternative scores, we weighted different parts of the exercise in different ways to see whether this would explain the ranking of trainees. In many tasks not all the component steps or items are equally important. Thus, it is not much good checking oil, tyres, battery and brake fluid before a desert drive if you don't also fill the tank. However, although there may have been some weighting by supervisors, we did not find any cases where apparently unclear or inconsistent criteria of judgement could be made fully consistent by a given weighting of items.

Table 2 summarises the results for a given exercise; that is, it does not show whether supervisors used the same standard for both assessment exercises, but only whether they seemed to be using the same criterion on all occasions when they used a particular test. The table distinguishes between each of the subgroups of supervisors and trainees involved in the study. 7 Column 1 shows the number of supervisors in each group who actually informed the researchers of explicit criteria which they were using, and column 2 the number whose rankings were consistent with underlying consistent criteria. Column 3 shows the percentage of trainees in that part of the sample who were tested by 'inconsistent' supervisors: a figure which ranges from 5% for supervisors' own micrometer exercises to 44% for their own invoicing ones.

TABLE 2 ABOUT HERE

In interpreting these results, it is important to note that the situations in which we assessed the supervisors' consistency were themselves not strictly alike. Supervisors are often dealing with only small numbers of trainees, whose performances in turn represent only a few of those possible. It was thus difficult to judge how consistently they dealt or would deal with the range as a whole - and especially with the difficult middle ground between those judged clearly competent and those judged clearly incompetent. In general, the invoicing results showed much greater variability, while some micrometer supervisors had only trainees with perfect or near perfect scores.

This said, it is apparent that there were significant differences in supervisor consistency which were related to the task (or occupational area) involved. Supervisors involved in micrometer testing were very rarely inconsistent, whereas in the invoicing subsamples about a quarter of the supervisors made one or more judgements which could not be reconciled with the criteria being used when assessing other trainees. The difference between occupational areas is itself probably related to the intrinsic nature of the task, and the ease with which explicit criteria can be devised, as well as to the patterns of trainee performance which the supervisors found. There seems no reason to treat one or other task as representative of workplace assessment as a whole, which in turn makes it impossible to set any overall figure for likely 'intra-assessor' consistency. Once again, we would expect the situation in school subjects to be comparable.

Although these figures indicate that, in some occupational areas, a considerable percentage of trainees might be tested by 'inconsistent' supervisors, it does not follow that they would be affected directly. On the contrary, it would appear from the research data that only a very few will receive assessments from their supervisors which differ from those they would receive if he or she were always consistent. This is because supervisors were not usually 'consistently inconsistent', in the sense that there was no apparent pattern at all to their judgements. It was rather that a few judgements were inexplicable. For example, one invoicing supervisor said that a trainee with scores of 16 items correct out of a possible 17 (raw and alternative) "could not" do invoices, and another with 15 out of 17 (raw) and 16 out of 17 (alternative) "could". At another scheme, all trainees with scores of 9 or more items correct (raw) or of 11 or more correct (alternative) were given a "can" judgement except for one, who with scores of 11 (raw) and 15 (alternative) got only a "maybe can".

On the other hand, it is also quite possible that our figure for the number of 'consistent' supervisors overstates the case. 8 Any supervisor who gives everyone a 'can' is, obviously, judged to be operating a consistent standard. So are a number who graded a small number of trainees for us, and had, say, 3 with very few errors ('can') and 1 with very many ('cannot'). We cannot know how consistently they would have operated in the middle range.

(2) During assessments, do supervisors judge a trainee's current performance on its own merits?

A possibility in any locally run system where instructors are also the assessors is that, in some cases at least, the results recorded will reflect the assessors' preconceptions rather than actual performance. Personal relationships between instructor and trainee may make the former more or less likely to give a positive report. So may preconceptions on the instructor's part about the sort of trainee who 'ought' to be able to complete a task, or who 'can't be expected' to succeed.9 Continuous assessment will not necessarily affect the situation either way. We all recognise, more or less consciously, that behaviour varies over time, and trainees' achievements may be classified as, say 'luck', or 'a chance slip'.

As noted above, all the supervisors involved in the study were asked for a preassessment of their trainees' competence on the relevant task. This was based on their previous experience and/or assessment of the trainees. Supervisors were sometimes reluctant to give this - perhaps for fear of being proved to be wrong. At other times, they felt they simply did not know the answer, because of the nature of the training, the number of trainees they dealt with, or their limited contact with the trainee in question. However, we were able to collect 'preassessments' for most trainees in the sample. The supervisors' preassessments were compared with their assessments of trainees' performance on the two test exercises. This gave us some indication of whether preconceptions affected the way supervisors judged their trainees in a relatively formal assessment context. *10*

Once again, substantive differences between the two task areas are important in interpreting the results. The micrometer sample in the 'alternate forms' part of the study (involving two exercises written by the research team) consisted largely of trainees who were preassessed as competent, and scored in the upper range of possible performance. One would not, from their actual performance, expect much divergence between pre- and post- assessment. Nor does one find it. Most are assessed as competent on both the exercises completed for the study; and only about 10% received different assessments before testing compared to one or other test exercise.

By contrast, the performance of the 'alternate forms' invoicing sample was very variable. If supervisors' assessments were based solely on actual performance at the time of assessment, one would expect considerable changes in judgement. This was indeed what we found. On the first exercise, 37% received an assessment of competence different from the preassessment; and on the second, 34%. *11*

In the second phase of the research - where fewer micrometer trainees were near the end of their year - there was less difference between task areas in the characteristics of score distributions. On the moderating exercise devised by the researchers, 20% of the micrometer trainees received a judgement different from their preassessment; and on their own supervisors' exercises, 18%. For invoice trainees, the corresponding figures are 43% and 42%.

Because supervisors always gave trainees two assessment exercises, we were also able to compare their judgements of performance on these two occasions. As discussed below, there was often very considerable variation in trainees' performance on the two occasions. If supervisors judged each occasion on its own merits, one would expect substantial numbers of trainees also to receive different judgements.

TABLE 3 ABOUT HERE

Table 3 summarises supervisors' behaviour in the different parts of the study. Comparison between exercises is most straightforward for the 'alternate forms' data, where the exercises were identical except for variation in the actual numbers used. As noted earlier, the performance of micrometer trainees, especially in the alternate forms study, was far less variable than that of invoicing trainees, both within and across exercises. The lower figure for supervisors giving, and trainees receiving, different judgements is consistent with this.

In general, the results are consistent with considerable ability and willingness on the part of supervisors to judge on the basis of the immediate evidence. Thus, a large majority of invoicing supervisors in the alternate forms study gave trainees differing assessments. The 'own exercise' data must be interpreted somewhat more cautiously, for though both exercises were again designed to test the same skill, they were often very different in style and content. *12* However, they also indicate that preconceptions about 'general' competence, or personal relationships with trainees, are not of major importance in determining supervisors' behaviour.

THE USE OF COMMON STANDARDS

Current assessment policy in the United Kingdom incorporates a strong commitment to 'national standards'. This is true of the Manpower Services Commission, for example, in its work on vocational qualifications. The Scots refer to national standards underpinning their new examination system, and seem to imply that these can be maintained straightforwardly enough through monitors and assessors. The English examination boards have always claimed that their procedures embody substantive standards, and are not merely concerned with establishing ordinal categories; but, under government directives, are now developing criterion referenced grade requirements (known as grade criteria).

Even with centralised question setting, small numbers of people and considerable continuity in personnel, such evidence as we possess suggests that examination boards' requirements and standards are far from absolutely fixed either between boards or over time. *13* Workplace supervisors have no examiners' meetings, and no yearly exam papers to define requirements for them. Occupational requirements and prectices will often change much faster than conventional academic school subjects. One way of providing for national standards in such a context is via centrally defined tasks or modules. However, as noted above, a reluctance to follow written instructions is a major stumbling block here, at least in the short term.

Alternatively a 'common occupational culture' may define standards already. Workplace supervisors may share standards derived either from the job's intrinsic demands, or from common training and experience. Indeed, other countries (such as Germany) seem to rely on such a shared unexplicated culture in so far as they take clear note of the standards issue at all. *14* In fact, the most important requirement for valid assessment is that we have a very clear idea of what we are trying to assess, and here specific workplace training would seem to be at a clear advantage *vis a vis* most school and even college programmes.

(1) Do they use a common criterion?

One aim of this research was to see both how far an apparently clear and precise task definition was translated into comparable exercises across site, and how far supervisors' behaviour reflected a common standard of judgement. Here, the more important - because more realistic - data are those from the "own exercise" substudy, where training schemes used an assessment exercise of their own.

As described earlier, the exercises varied enormously in content and length (see table 1). Such differences do not, in themselves, preclude the exercises being tests of the 'same' skill, and being applied using some unambiguous (if rarely specified) standard. *15* Supervisors who wrote their own exercises also all used our own 'moderating' exercises. Consequently, if the relative performance (or rank) of trainees at a given site had been the same on both exercises, we could have reached some tentative conclusions about whether individual supervisors seemed to be applying a similar success criterion on both our and their exercises. We could also have looked at whether supervisors on different sites were using standards for success in their own exercises which stood in the same relation to levels of performance on our moderating test. However, the performance of trainees who completed an 'own site' and a moderating exercise was in fact highly variable. *16* We could therefore not use the moderating exercise to deduce anything about cross site standards on the sorts of exercises workplaces may set themselves. *17*

What was possible was to compare the criteria which supervisors used in assessing performance on the researchers' exercises. Tables 4 and 5 show the criteria used by micrometer supervisors involved in the alternate forms study in terms of the total number of trainees given a particular classification. Tables 6 and 7 show those used by their invoicing counterparts. They plot supervisors' judgements against 'raw scores' - i.e. the number of individual items correct. A '1' signifies a judgement that the trainee can do the task, a '4' that they cannot, and '2' and '3' are intermediate. If supervisors were entirely consistent in their standards, there would never be more than one entry in each row of the graph, because a given raw score would always receive the same judgement. *18* The more scattered across the graph the entries are, and the larger the number of trainees in

'outlying areas', the less consistent the standards used.

Tables 4 and 5 show micrometer supervisors in this sample to be using similar criteria in judging success - although trainees' results are also clustered at the upper end of the distribution, and for those scoring 10 or under there is less agreement. Tables 6 and 7 show that scores associated with a given overall judgement are more dispersed for the invoice data. This is consistent with other indications of a commonly agreed on task definition for micrometer use, and indeed with the more closely shared craft background of engineering supervisors. However it is also probably, in part at least, a function of less clustered scores.

TABLES 4,5,6,7 ABOUT HERE

(2) Do they 'norm reference'?

The discussion so far has been in terms of particular criteria, and whether, for example, total accuracy, or particular totals of items correct, were acceptable evidence of competence. However, it is possible that what supervisors were doing was simply 'passing' a given percentage of their trainees - whether 50%, 75% or even 100%. (This would, in itself, generate apparently different 'standards'.)

To check this, we looked at the number and proportion of trainees whom supervisors judged competent on the moderating and on their own exercise. The combination of variable trainee behaviour (see below) and small samples means that supervisors operating with some sort of substantive standard will often arrive at very different 'pass rates' with the same group of trainees. This is indeed what happened. While we obviously cannot conclude from it that supervisors applied some constant Platonic standard, they also did not, for the most part, automatically 'pass' the top x%.

Overall, the results indicate considerable variation in the standards used by supervisors. This could affect a significant percentage of trainees - the more seriously because 'criterion referenced' tests generally (though not necessarily) only allow pass and fail. To use the analogy of the driving test yet again: different standards would mean that in one place people were free to drive almost anything on four wheels, who elsewhere would be unable to get from country to town for work because they had failed the test.

The results of the study also indicate, again, that there may be major inter-occupational differences. These may be related to whether a common training programme has been followed by most supervisors, since standards were more consistent in the micrometer sample. The 'intrinsic demands' of the invoicing task - that the customer receive an invoice which is entirely correct - seemed by contrast to be a less powerful influence. *A priori*, there seems no reason to suppose that one or other occupational area is the more 'representative'. Industry statements (implying that 100% accuracy is the general standard) would have led one to expect greater agreement, in this case, among invoicing than among micrometer supervisors. It thus also seems

impossible to predict, in advance, what the situation will be in any particular field.

ASSESSMENTS AS PREDICTORS OF PERFORMANCE

Practical assessments have the advantage over more conventional tests of high 'face validity'. In other words, their relationship to the behaviour one is interested in tends to be direct and obvious. Such evidence as we have on test validity is largely American, and in fact the largest studies of tests' predictive validity have been carried out by the US Armed Forces. They confirm that assessments will provide better predictions of the future behaviour in which we are interested, the closer they come to replicating situations in which it occurs. This is hardly surprising, but it is nonetheless the main - and some would say the overwhelming - argument for attempts to move away from traditional assessment procedures.

It does not follow, however, that any and all practical assessments are automatically a 'good thing', whose results can be regarded as some form of absolute truth. A score on a 'practical' test or sample of work behaviour is, just as much as any other test score, based on a sample of behaviours, from which generalisations are made. At the same time, it is subject to other uncontrolled variation. On any given occasion all sorts of chance errors may affect performance in either direction.

Much of the empirical work on public examinations has concentrated on intra- and inter- marker reliability - comparable to the investigation of supervisors' judgements described in the previous section. Use of 'parallel' or 'alternate' forms of a test with the same subjects is also a way of looking at the test's reliability when used with a particular group of candidates.*19* It is also especially relevant in the context of competency testing, and especially decentralised practical competency testing, because an 'alternate forms' method of assessment is, in effect what will operate. In, for example, a written public examination - GCE in England, or SATs in the United States - one might use alternate forms to estimate the reliability of a test which will be taken by all candidates in a single, fixed form. In much vocational testing, and in the teacher assessed parts of school exams, the behaviour to be assessed will be defined, but the exact content and format of the test probably will not. A large number of 'alternate forms' will actually operate; and the more pairs can be examined, the better one can place likely bounds on the reliability of the system once it is in operation. *20*

In this sense, the whole of the current study can be seen as involving 'alternate forms' of a particular assessment. However, the main source of information on reliability was the first 'alternate forms' substudy or phase of the research. Here, as described earlier, the research team designed all the tests which supervisors administered, and all were, as far as we could determine by inspection, identical in form except that the numbers were varied systematically. In administering these tests, we also were able to vary systematically

the order in which trainees completed them. Reliability estimates should not, therefore, be affected by learning or by systematic 'boredom' effects.

The reliability estimates obtained are also of more general interest because of the increase in tests which, like these, are criterion referenced. In this context a potential problem of which we were aware is that it is extremely difficult to separate the differences between the test reliability and the difficulty of supposedly 'parallel' tests. Thus we may be thought to have produced intra-subject variation by definition, through our variation of the actual numbers used in otherwise identical tests. However, while this may matter if one is interested primarily in the reliability of one of the tests examined - since it alone is to be used generally - in the context of assessments which are set and marked locally, it is the variability between alternate forms which is of relevance. In other words, the variation we wish to examine is the sum of measurement error and variation between different (alternate) tests. *21*

We may also add that results suggest that, in this context, differences in test difficulty resulting from differences in actual numbers used may be a relatively minor concern. One set of criteria which has been used to decide whether tests are essentially parallel relates to whether they in fact produce the same sort of distribution of scores within sampling error. In analysing the data we looked at the various test forms' distributions from this point of view. The overall means and variances for the various forms did not indicate any clear differences in underlying distributions.

TABLES 8 AND 9 ABOUT HERE

Tables 8 and 9 summarise the reliability estimates obtained for the exercises (see appendix I for the formulae employed). In each case, we have given three estimates. One uses an estimate of error variance derived from the tests taken first by the trainees in the group. One uses an estimate derived from the tests taken second. The third (which, by definition, falls between the two) is the population coefficient of correlation between trainees' scores on the two tests. All estimates are on the basis of the number of items answered correctly. No 'pass-fail' or other cut offs were used.

Other things being equal, one might have expected lower reliability estimates for the engineering and mechanics trainees who completed the micrometer exercises. Tests will tend to be more reliable the more varied the 'true scores' of the population, simply because it is the less likely that a slight change in an observed score will involve a change in the candidate's rank. The micrometer sample's scores were heavily clustered at the 'high' end of the distribution, probably because most had almost completed a year's training.

Given this homogeneity of the 'alternate forms' micrometer sample, the reliability estimates shown in Table 8 seem reasonably satisfactory. Only those for the AE (metric) pair are consistently low, and at this level of analysis the sample size is small. This would indicate that,

on our exercises, and for this population, one could generalise quite reliably from trainees' performance. However, as we mentioned earlier, many of the schemes' own exercises differed considerably from the research team's. Because all such exercises contained measurement - and many consisted only of this - we recalculated inter-test reliability using the measurement items only. It is interesting that these estimates were generally as high as or higher than those for the exercises as a whole.

The reliability estimates for the invoice data are, by contrast, mostly very low. Looking at Table 10, we find a considerable number of cases where estimated reliability is zero, and many others where the relationship between scores on the two exercises is clearly very weak. Even for the whole sample, which is quite large, none of the estimates approach the levels which are normally considered acceptable for testing purposes.

These results are particularly disturbing if we compare them with those obtained for internal reliability. Many investigations of reliability do not obtain duplicate measurements for the sample of candidates studied, but instead focus on the internal consistency of the test: whether different questions, or different combinations of questions, tend to produce the same results as each other. One way of doing this is to split the test in half, and look at the agreement between the two halves. However, in this case, it seemed more appropriate to use calculations of 'coefficient alpha' which is, in effect, the average value of all the estimates one would get if one took every possible way of dividing the test into two halves. This was because the basic 'split half' approach is only really applicable when the test can reasonably be seen as falling into halves which test the same behaviour(s). *22* In this study, this is arguable for the invoice tests, but less so for the micrometer ones, where we intentionally broke out ability to set the instrument; understanding of tolerances; and accurate measurement of items. *23*

Tables 10 and 11 present internal reliability estimates calculated using coefficient alpha only, and so allow comparison between the two skill areas. In addition, reliabilities for the invoice exercises (Table 9) are presented separately for the two invoices involved in each.

TABLES 10 AND 11 ABOUT HERE

The difference between the two estimates is especially important because, in many studies of public examinations and tests, only internal reliability is studied. In this case internal reliability estimates might well have led to the view that a single invoice exercise was a trustworthy indication of trainee capacity. Such a conclusion would be highly misleading. Performance on one invoice exercise in fact tells one virtually nothing about how a trainee will perform on another.

The conclusion to be drawn is not, we believe, that the invoicing exercises were somehow faulty, and needed to be 'tightened up' or

redrafted. Rather, we are observing extremely variable behaviour on the part of the trainees. The invoice exercises, like the micrometer ones, were not single item tests, but simulations each of which involved two invoices. This makes their high internal reliability unsurprising. We deduce that, among micrometer trainees, most had mastered the skills involved to the point where they were secure. Invoice trainees, by contrast, were often still in the process of mastering the procedues and skills required. Theirs was a 'soft' competence, so that for many it would be untrue to say that they could not do invoices - but also highly unwise to rely on their doing so correctly. Such a state of affairs is likely to be common in any educational or training context.

The relationship between test performance and supervisors' preassessments underlines how difficult, and often misleading, it can be to assess the 'competence' of learners. 24 The reliability results for the invoice data point up very strongly the dangers of a single assessment. What is more, these results are based for the most part on very closely related tests administered while under observation. Additional confirmation of the variability of trainee behaviour comes from those sites where we were able to make some comparison between our tests and those designed by supervisors, following our specifications (more or less).

As explained above, the very different content of these tests meant that we could not compare scores directly. However, where there were enough trainees on a given site, we could look at rank correlations.

TABLES 12 AND 13 ABOUT HERE

Results are shown in Tables 12 and 13 and confirm that the relation between trainees' performances on two exercises supposedly concerned with the same behaviour may be extremely low. Once again, the pattern for micrometer use is markedly different, with rank correlations as high as 0.9 - but also as low as -0.2! For invoicing, the relationship appears often to be effectively non-existent.

Correlations as low as these, and the underlying variability of behaviour they reflect, must be of concern to all of us involved with the growth of criterion referenced or competency testing: and the more so, the higher the stakes involved. The answer, however, surely is not to fall back on the misleading textbook counterpoise between reliability and validity, and argue that we must, after all, sacrifice the latter to the former. We must go on assessing complex activities in applied situations because it is to this type of behaviour that we wish to generalise. What we do need to rethink is our dependence on single tests to do so; and our apparent addiction to "pass-fail" criteria which are taken to represent the possession or otherwise of some general skill.

FOOTNOTES

1. A more detailed account is given in Appendix I.

2. Invoices were scored in two ways. "Raw scores" show number of items correct. Under "alternative" scoring, an item which is wrong only because it incorporates a previous error is not counted as wrong. (For example, if a trainee calculated a discount incorrectly, their overall charge to the customer was affected. However, if they subtracted the discount - rather than adding it! - and if their final total was correct in terms of the amounts they were using, alternative scoring would only count one error. The trainee would not be penalised for an incorrect discount and an incorrect total.)

3. i.e. the schemes in the "own exercise" substudy. In the "alternate forms" part of the research, only researchers' own exercises were used.

4. It was impossible to know exactly how many supervisors used the 100% criterion because of the score distributions. If a supervisor judged all invoices inadequate but they were all well short of 100% accurate, one could not know how they would actually judge one close to 100%. Similarly, one could not know the actual criterion that would be used by a supervisor who had, say, one or two trainees with 100% accurate invoices (judged as satisfactory), and others with invoices which were highly inaccurate (and judged as unsatisfactory).

5. See Gipps C, Steadman S, Blackstone T and Stierer B, Testing Children, London: Heinemann Educational Books, 1983

6. Examples are taken from the Scottish Examination Board's grade related criteria for English.

7. i.e. those administering two alternative forms of the researchers' exercise, and those administering their own and the moderating one. Each of these groups is further subdivided between micrometer and invoicing samples.

8. We judged them 'consistent' if they appeared so on one set of marks.

9. Another potentially serious problem is that trainees' success rate reflects on the instructor. However, this is likely to affect the criterion or standard used for the group as a whole rather than judgement of individual trainees.

10. One cannot safely generalise from this to the way a supervisor conducts and modifies continuous assessments.

11. 33% if one does not distinguish between 'maybe can' and 'maybe cannot'.

12. See table 1 above.

FOOTNOTES

13. See, for example, Nuttall DL, Backhouse JK & Wilmott AS, "Comparability of Standards between Subjects" (Schools Council Examinations Bulletin 29: Evans Methuen Educational 1974), and Christie T & Forrest GM, Defining Public Examination Standards (Schools Council Research Studies: Macmillan Education 1981)

14. See Wood R, "Assessing Achievements to Standards" (unpublished paper prepared for MSC Quality Branch, 1985)

15. Though they obviously make it less likely.

16. With negative rank correlations in a number of schemes. See tables 10 and 11 below.

17. This follows necessarily from the combination of varied assessment exercises and variable trainee behaviour. The impossibility of comparing supervisors' standards directly, independently of a moderating exercise, is not something we could have got around by changing the research design.

18. This assumes no weighting of particular items. Although such weighting would seem quite reasonable, we could not discern any clear patterns of weighting in the supervisors' given assessments.

19. By examining the degree of divergence in scores, one can estimate how far the assessment will, indeed, provide consistent results. A large enough number of strictly parallel tests would, by definition, produce the same result as a large number of repetitions of the same test. The strict definition of parallel forms is that:

    (1) the tests overlap completely in their true-score distributions;
    (2) variance of errors of measurement on one form is the same as on the other;
    (3) true scores covary with errors of measurement
    (4) errors of measurement covary zero from one form to another.

    However, in the vast majority of actual test situations, at least in the United Kingdom, assessors operate using alternate forms of a test which are judged, on the basis of experts' face examination, to be "close enough".

20. Cronbach LJ, Gleser GC, Nanda H and Rajaratnam N, The Dependability of Behavioural Measurements (New York: Wiley, 1972) subsume reliability within validity as an aspect of generalisation. Obviously in examining reliability, one selects a group of candidates comparable to the population expected to take the test. Thus, we gave the invoicing exercises to YTS trainees following clerical courses and with some experience with invoices, not to all clerical trainees, nor only to those attempting the difficult BEC National Certificate.

FOOTNOTES

21. However, it is always possible that underlying differences in true score distributions will be masked by the error score distributions (or vice versa). An alternative method, which would involve comparing the relationship between each of the supposedly parallel tests and another independent variable was rejected as too expensive and time consuming.

22. i.e. when the items are unidimensional or test the same behaviour.

23. Because there are several questions of each type - 4, 6 and 4 respectively - many possible 'halves' compared in coefficient alpha calculations will include questions of each type.

24. It should be emphasised that preassessments were just about as weakly - or strongly - related to exercise performance as the two performances were to each other.

APPENDIX I

The data analysed here were all collected during the period May to November 1984, and describe the performance of Youth Training Scheme (YTS) trainees on practical exercises designed to test their ability

either to use a micrometer

or to complete an invoice.

Any trainee completing an exercise had received training and/or experience in that area, although this obviously varied in quantity (and, presumably, quality). All exercises were administered by the trainees' supervisors, who also were asked to provide

1 a preassessment of the trainees' ability

2 a judgement of their competence as displayed by the exercise in question.

All trainees were asked to complete two exercises. This gave three opportunities per trainee for the supervisor to provide an overall judgement. The overall judgement was in the form

can/cannot (complete invoice forms successfully / use the micrometer successfully

Supervisors were observed during testing by one of the research team, and their method of presentation was recorded.

The exercises used were all designed to be in line with the type of assessment vehicle currently seen by the British government as appropriate for YTS assessment. All sites provided an opportunity to observe supervisors' methods of assessment, and consistency of judgement. However, the research design also provided for two quite distinct phases of data collection, and the analyses are for the most part concerned with one or the other.

The first, which we refer to as the 'alternate forms' study, involved trainees completing two exercises, both of which were written by the research team. They were designed to be 'alternate forms' of a micrometer or invoice test respectively, and varied only in the numbers used. (See below) This was because we were concerned to see how far the use of different numbers may affect, in systematic ways, the difficulty of what are apparently very similar tests.

A larger number of tests used in piloting was reduced, for data collection purposes, to 6 micrometer exercises (4 metric, 2 imperial) and 4 invoice exercises. The imperial exercises were given whenever they were the form familiar to trainees; but the resulting sample size was very small.

Each set of 4 was conceived of as two pairs: that is, the pairs varied systematically from each other in certain ways, but without our knowing in advance whether one pair would be 'more difficult'. In the case of the micrometer exercises, A and C involved a setting of the '3.01' and '5.09' type, which we knew from previous work to be confused frequently with '3.1' and '5.9'. E and G both had settings requiring two turns of the barrel beyond the millimetre unit (eg '10.79').

Trainees doing test A also did E or G; so did those doing C (giving us four 'metric' pairs: AE, AG, CE, CG). Test order was also varied within site. Each 'type' of question was asked more than once. Thus, on each exercise, trainees were asked to set the micrometer to 4 different readings; to answer 6 questions on tolerances; and make 4 actual measurements. (Most previous research on the effect of numbers in tests deals with single item comparisons. With a larger number of items covering a topic, the impact of particular numbers may be reduced.)

In the case of the invoice exercises, the exercises were again divided into two pairs. One pair required the use of a catalogue, while the other had the unit prices included against the items requested. All the invoice exercises included an invoice with a discount calculation and one without, where the trainee had to decide not to give a discount.

Trainees did C or D, and E or F (giving 4 'pairs' again: CE, CF, DE, DF) and order was varied within site. Analysis of the results did not show any systematic differences in the difficulty of the tests, but the results of the 'alternate forms' part of the study are our major source of information on test reliability - or the consistency of trainee performance.

In the second part of the study, trainees did one exercise prepared by the researchers (the 'moderating' exercise), and one prepared by their own scheme. (This part of the study is referred to for short as the 'own exercise' substudy.) Each site was sent a summary of procedures to be followed in designing an exercise, and sites involved were subdivided into those which were given a very simple description of the behaviour to be tested, and those given a more detailed one, modelled on the 'Standard Task' format being developed by the Manpower Services Commission, the agency responsible for youth training. These are reproduced below.

A: BASIC SUMMARY OF PROCEDURES

(1) Trainees are asked to complete two exercises, each of which is intended to determine whether they can

(a) complete invoice forms correctly, or
(b) use a micrometer.

One exercise is designed by the instructors; the other is a prewritten exercise used by the researchers at all sites. The

latter is self explanatory.

(2) Both exercises are administered by the instructors. Beforehand, they are asked whether or not, in their opinion, the trainees possess the skill involved.

(3) On the basis of each trainee's performance on each exercise, the instructor is asked whether he/she would judge that the trainee had completed the exercise competently - ie a simple can/cannot judgement.

(4) The researchers will take away with them, in addition to their records of the instructors' judgements:

(a) a record of the trainees' answers, and

(b) a record of the exercise devised by the instructors.

(5) All schemes instructors and trainees will remain anonymous. We would like to stress to all participants in the research that it is the method of assessment which is being tested - not the people.

Standard Task Format: Use of Micrometer

Trainees complete two exercises, each of which is intended to determine whether they can use a micrometer correctly. One exercise is designed by the supervisor or training officer responsible for the trainees. The other is a prewritten exercise used by the researchers at all sites.

Both exercises are administered by the supervisor in charge, who will also be asked for his or her judgement of the trainees' capabilities prior to their completing the exercises. One of our project team will be present to observe.

We need to be able to take away the trainees' responses together with a copy of your exercise. We will be happy to return your test results to you if you so require.

Task Definition

When designing your micrometer exercise, please ensure that it tests

(a) ability to make an accurate measurement;

(b) understanding and application of tolerances.

Criterion for success:

Measurement accurate to 0.01mm or measurement accurate to 1 thou

Standard Task Format: Completion of Invoices

Trainees complete two exercises, each of which is intended to determine whether they can complete an invoice form successfully. One exercise is designed by the supervisor or training officer responsible for the trainees. The other is a prewritten exercise used by the researchers at all sites.

Both exercises are administered by the supervisor in charge, who will also be asked for his or her judgement of the trainees' capabilities prior to their completing the exercises. One of our project team will be present to observe.

We need to be able to take away the trainees' responses together with a copy of your exercise. We will be happy to return your test results to you if you so require.

Task Definition

When designing your invoice exercise, please ensure that it includes

(a) more than one invoice;

(b) percentage calculations;

(c) addition and multiplication.

Criteria for success:

(a) invoice filled in correctly;

(b) calculations correct.

As in the first phase of the study, the trainees involved completed both exercises (the moderating exercise and their own scheme's) on the same day. The trainees' completed responses were collected, as well as the supervisors' judgements and examples of the exercises themselves. All scoring by the research team was on an item by item (or calculation by calculation) basis. This meant that the research team's micrometer exercises had 14 items, and their invoice exercises 17. Schemes' own exercises varied greatly in the number of items included.

Formulae for Reliability Coefficients

We looked at parallel forms reliability and the formula used was:

$$\rho_{ff'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

where $\rho_{ff'}$ is the population coefficient of correlation between scores on parallel forms f and f'.

σ_x^2 is the population variance of the observed scores.

σ_e^2 is the variance associated with errors of measurement.

Coefficient alpha is for assessing the internal reliability of the test. The formula is given by:

$$\frac{I}{I-1}\left[1 - \frac{\Sigma p_i q_i}{S^2 \quad (\sum_{i=1}^{I} X_i)}\right]$$

where there are I items in the test.

p_i is the proportion of people who gave a correct answer to item i and $q_i = (1-p_i)$. The S^2 term refers to the variance of the total scores on the test.

Table 1

CONTENT ANALYSIS OF SCHEMES' OWN INVOICING EXERCISES

Number of tasks	No of supervisors
1	3
2	6
3 or more	10

Number of items in task	No of supervisors	
	Task 1	Task 2
1	1	2
2-4	2	2
5-7	6	1
8-10	8	9
over 10	2	2

Type of task*	Task 1	Task 2 or above
Invoices:		
Filling in Checking	15) 2)	14
Customer delivery entries	1	-
Pay at end dockets	1	1
Pay at end card	-	1
Transferring and entering numbers	-	1
Wage calculations	-	1

* Where more than two tasks were used, all tasks after the first are included

CONTENT ANALYSIS OF SCHEMES' OWN MICROMETER EXERCISES

Number of tasks	No of supervisors
1	14
2	6

Type of task	No of supervisors	
	Task 1	Task 2
Measurement only	16	-
Measurement and manipulation	1	-
Setting only	-	1
Written test	1	-
Measurement and theoretical questions	2	-
Measurement and understanding tolerances	-	3
Verbal questions	-	1
Vernier	-	1

TABLE 2 : CONSISTENCY OF JUDGEMENT CRITERIA

Type of exercise (N = numbers of supervisors involved)	Number of Supervisors who Stated explicit criteria	Number of supervisors using Consistent criteria - stated or apparent - on raw and/or alternative scores	Number and % of trainees tested by "inconsistent" supervisors
Micrometer Exercises: Alternate Forms N = 31	-	28	N = 17 % = 13
Invoices: Alternate Forms N = 27 (actual number 28 but 1 supervisor only graded 1 of his 2 trainees)	-	21	N = 52 % = 40
Micrometer: Ours-Theirs Moderating exercise (N = 20)	-	19	N = 5 % = 6
Micrometer: Ours-Theirs Their exercises N = 20	4	19	N = 4 % = 5
Invoicing: Ours-Theirs Moderating Exercises N = 19	-	15	N = 35 % = 25
Invoicing: Ours-Theirs Their exercises N = 19	2	13	N = 60 % = 44

* These figures show consistency within a given exercise, not across exercises.

Table 3

CHANGES IN JUDGEMENTS OF TRAINEES' PERFORMANCES BETWEEN DIFFERENT EXERCISES

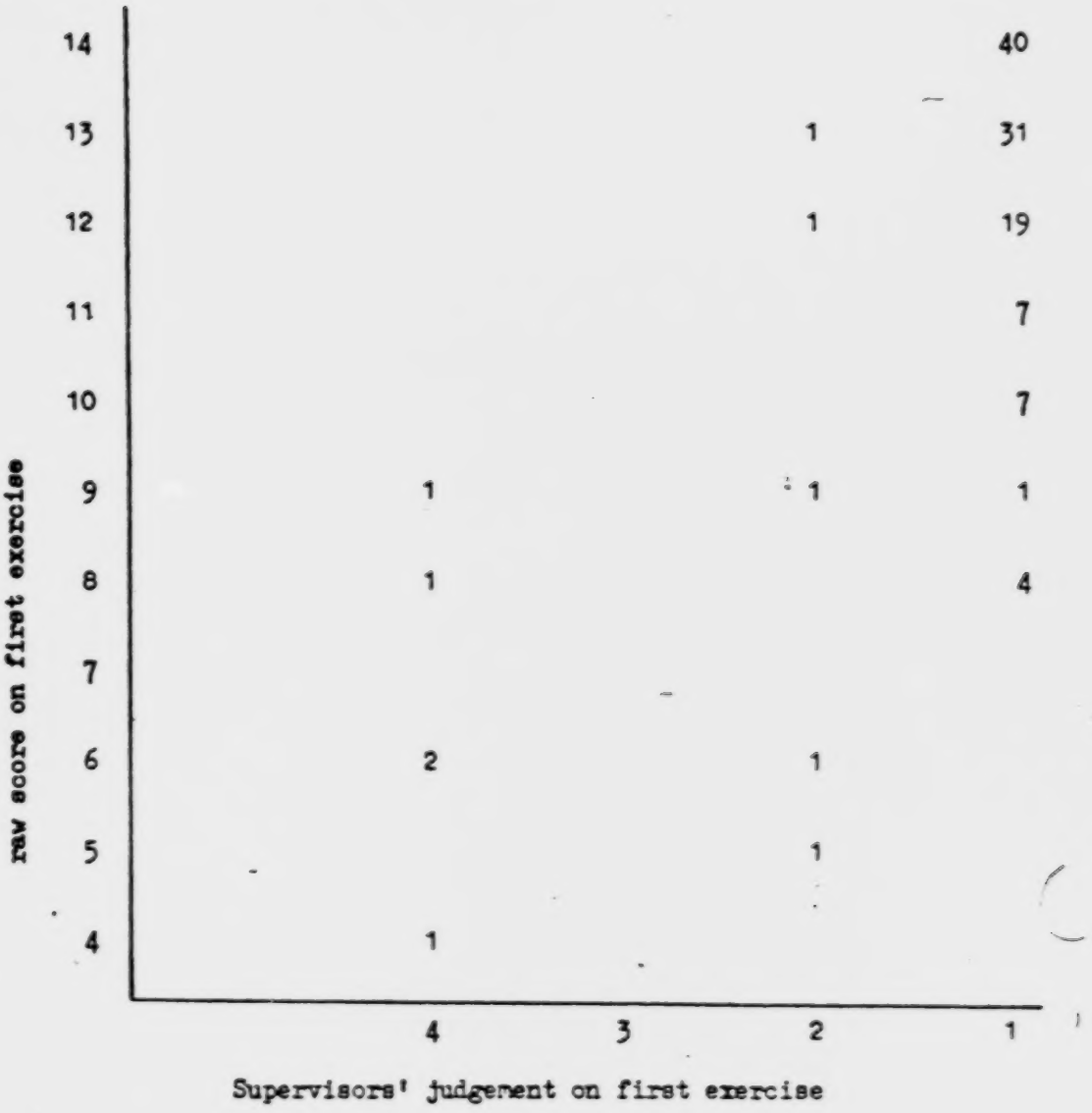
	Total no of supervisors	Percentage of supervisors giving different judgements on one or more trainees * (N = number so doing)	Percentage of trainees receiving different judgements * (N = number so doing)
Micrometer Phase I (Alternate forms)	31	16 (N = 5)	8 (N = 9)
Invoicing Phase I (Alternate forms)	28 (26*)	73 (N = 19)	35 (N = 51)
Micrometer Phase II ('ours/theirs')	20 (19*)	53 (N = 10)	21 (N = 15)
Invoicing Phase II ('ours/theirs')	19 (18*)	83 (N = 15)	45** (N= 50)

* Omitting missing data

** 50 trainees received only task by task judgements on their supervisors' own exercises, rather than an overall assessment. Of these, 58% (N=29) received different judgements on the first task from that received on the researchers' exercise.

Table 4

GRAPH SHOWING SUPERVISORS' JUDGEMENT ON FIRST EXERCISE AGAINST RAW SCORE ON FIRST EXERCISE FOR MICROMETER SAMPLE BY TRAINEE : "ALTERNATE FORMS"



9 missing cases

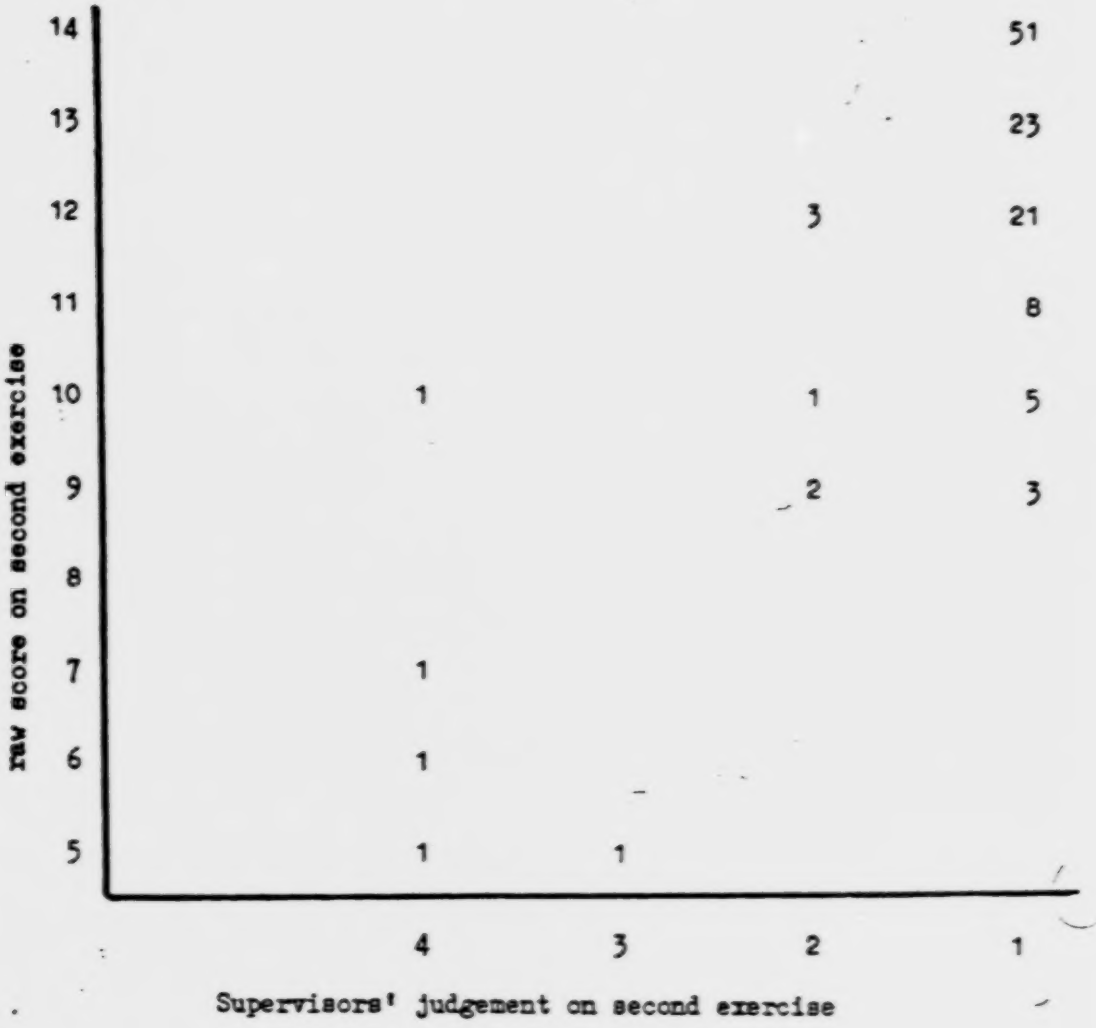
1 signifies a judgement that the trainee can use the micrometer successfully, 2 = probably can, 3 = probably cannot, and 4 = definitely cannot.

There were 14 items in the exercise, hence a maximum score of 14.

The numbers on the graph show the total number of trainees in the sample receiving a particular combination of raw score and judgement

Table 5

GRAPH SHOWING SUPERVISORS' JUDGEMENT ON SECOND EXERCISE AGAINST RAW SCORE ON SECOND EXERCISE FOR MICROMETER SAMPLE BY TRAINEE: "ALTERNATE FORMS"



6 missing cases

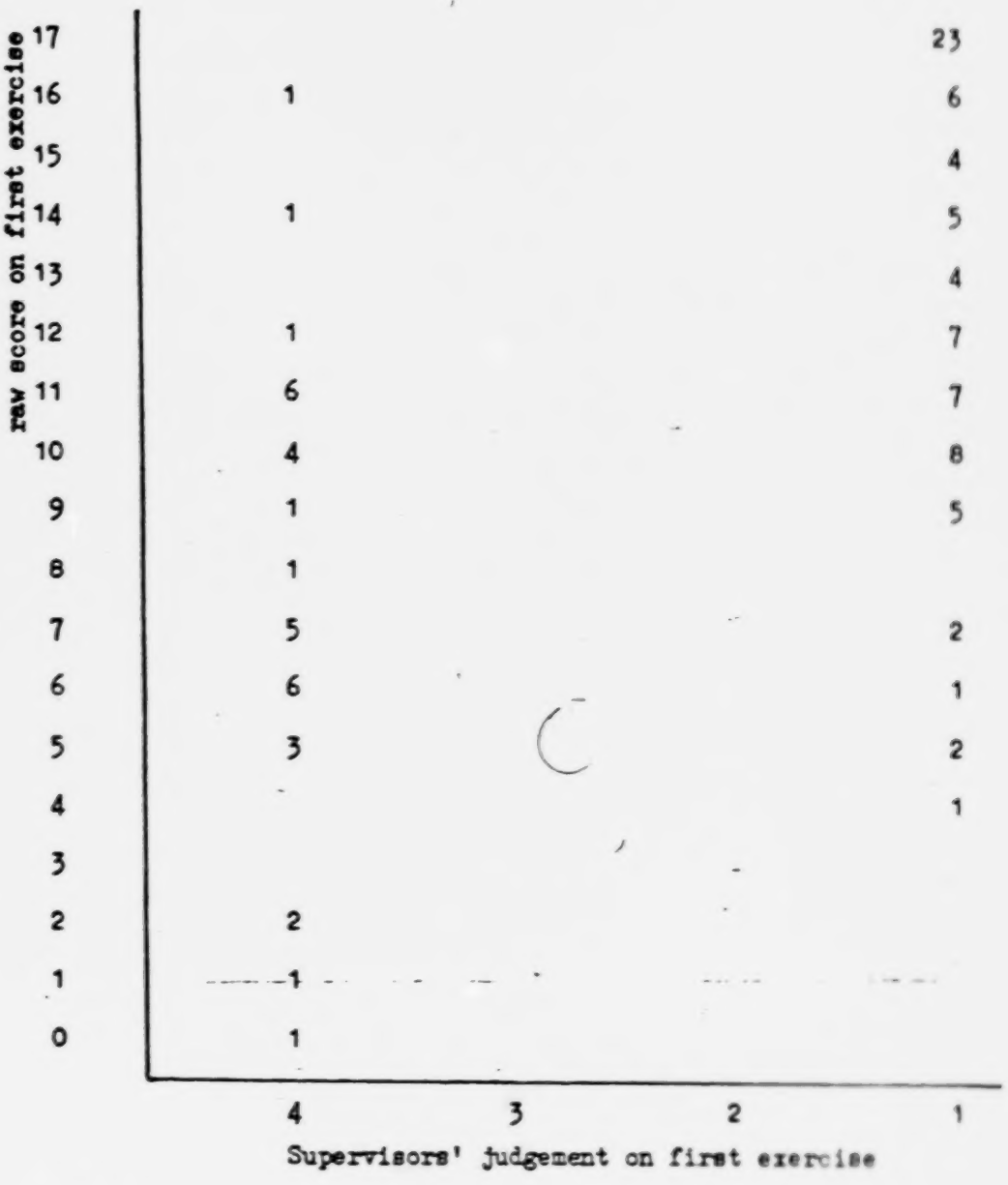
1 signifies a judgement that the trainee can use the micrometer successfully, 2 = probably can, 3 = probably cannot, and 4 = definitely cannot.

There were 14 items in the exercise, hence a maximum score of 14

The numbers on the graph show the total number of trainees in the sample receiving a particular combination of raw score and judgement.

Table 6

GRAPH SHOWING SUPERVISORS' JUDGEMENT ON FIRST EXERCISE AGAINST RAW SCORE ON FIRST EXERCISE FOR INVOICE SAMPLE BY TRAINEE: "ALTERNATE FORMS"



(11 missing cases)

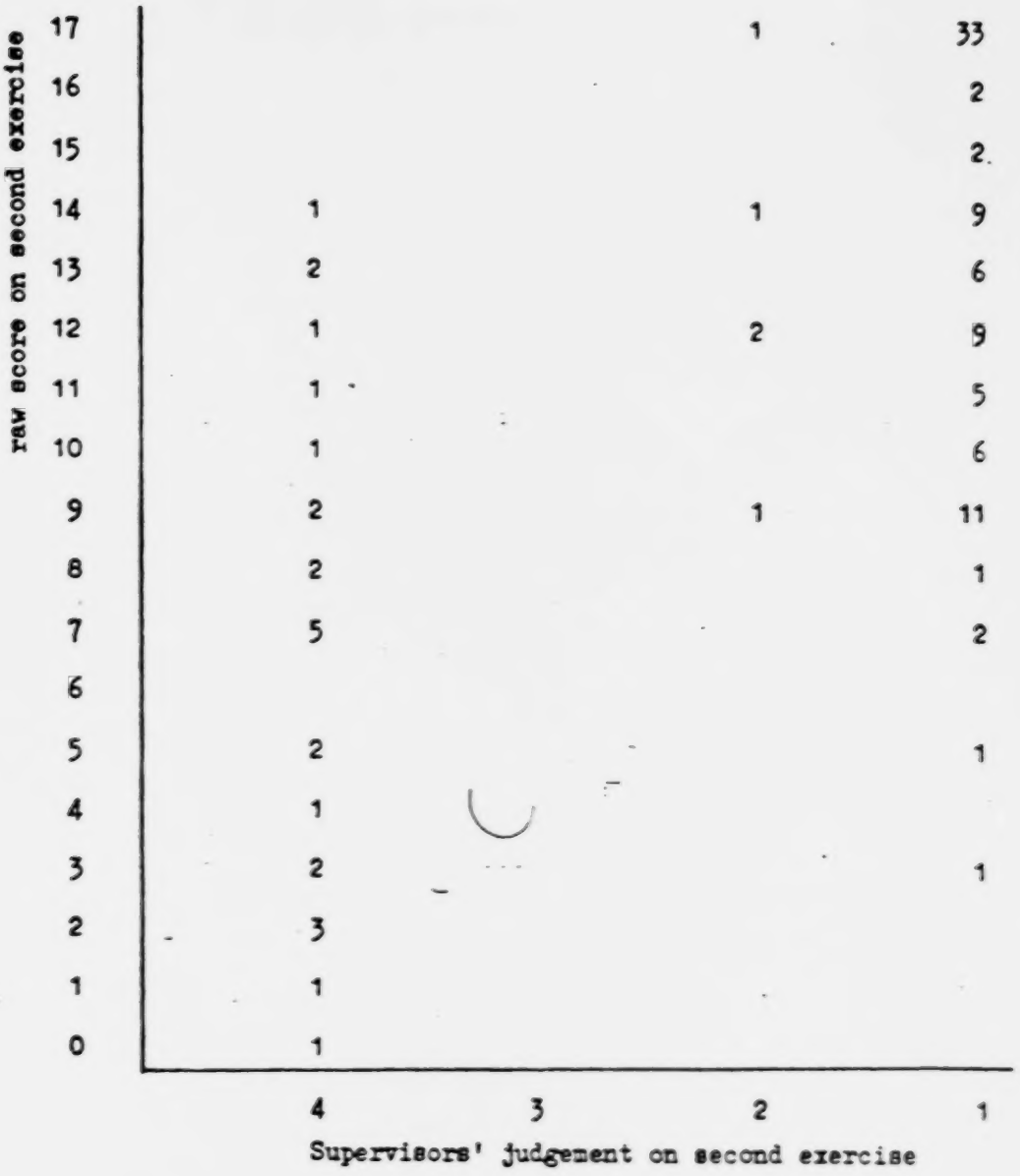
1 signifies a judgement that the trainee can complete invoice forms successful
2 = probably can, 3 = probably cannot and 4 = definitely cannot.

There were 17 items in the exercise, hence a maximum score of 17.

The numbers on the graph show the total number of trainees in the sample receiving a particular combination of raw score and judgement.

Table 7

GRAPH SHOWING SUPERVISORS' JUDGEMENT ON SECOND EXERCISE AGAINST RAW SCORE ON SECOND EXERCISE FOR INVOICE SAMPLE BY TRAINEE : "ALTERNATE FORMS"



(11 missing cases)

1 signifies a judgement that the trainee can complete invoice forms successfully, 2 = probably can, 3 = probably cannot, and 4 = definitely cannot.

There were 17 items in the exercise, hence a maximum score of 17.

The numbers on the graph show the total number of trainees in the sample receiving a particular combination of raw score and judgement.

Table 8

RELIABILITY ESTIMATES FOR MICROMETER SAMPLE (ALTERNATE FORMS)

	Raw scores
WHOLE SAMPLE (128)	
1st test	0.7285
2nd test	0.6912
	(Correlation: 0.7179)
AE (METRIC) PAIR (25)	
1st test	0.4323
2nd test	0.3011
	(Correlation: 0.3755)
AG (METRIC) PAIR (28)	
1st test	0.9080
2nd test	0.8971
	(Correlation: 0.9043)
CE (METRIC) PAIR (24)	
1st test	0.7673
2nd test	0.6078
	(Correlation: 0.7323)
CG (METRIC) PAIR (40)	
1st test	0.6161
2nd test	0.4828
	(Correlation: 0.5653)
BF (IMP.) PAIR (11)	
1st test	0.7366
2nd test	0.7864
	(Correlation: 0.7684)

Reliability Estimates for Invoice Sample (alternate forms)

	Raw Scores		Alternative Raw Scores*	
	inv. with discount	inv. without discount	inv. with discount	inv. without discount
WHOLE SAMPLE (130)				
1st test	0.1757	0.3757	0.5067	0.4991
2nd test	0.3205	0.5428	0.6652	0.6570
Correlation	0.2481	0.4593	0.5859	0.5781
**CE PAIR (41)				
1st test	0.3833	0.6734	0.7510	0.6302
2nd test	0.5004	0.7886	0.8025	0.7700
Correlation	0.4419	0.731	0.7767	0.7001
CF PAIR (25)				
1st test	0.0563	0.3263	0.1672	0.1867
2nd test	0.2292	0.5612	0.6722	0.6105
Correlation	0.1427	0.4437	0.4197	0.3986
DE PAIR (28)				
1st test	0.0000	0.0000	0.0000	0.0000
2nd test	0.1653	0.0857	0.6107	0.-193
Correlation	0.0827	0.0429	0.3053	0.2097
DF PAIR (36)				
1st test	0.3531	0.3327	0.6021	0.7543
2nd test	0.0000	0.1619	0.1157	0.4252
Correlation	0.1765	0.2473	0.3589	0.5897

*Alternative Raw Scores are derived by assuming the input to each question, and checking to see whether the answer is correct for that input.

** Invoice exercises were labelled "C", "D", "E", and "F" (see table 12). Each contained two invoices, and each trainee completed two exercises - or 4 invoices - in all. One exercise in each pair required use of a catalogue, and the other did not.

Table 10

Internal Reliability Estimates for Micrometer Sample (Coefficient α)

Whole sample	
1st test:	0.7678
2nd test:	0.7602
Exercise A:	0.7045
Exercise C:	0.7377
Exercise E:	0.7192
Exercise G:	0.7935

TABLE 11

Internal Reliability Estimates for Invoice Sample (Coefficient α)

(alternate forms)

Exercise	Raw Score	Alternative Raw Score *
C inv. with discount	0.9280	0.8526
inv. w/o discount	0.9157	0.8709
D inv. with discount	0.9349	0.6058
inv. w/o discount	0.8172	0.6058
E inv. with discount	0.9541	0.9369
inv. w/o discount	0.9643	0.9648
F inv. with discount	0.8721	0.6835
inv. w/o discount	0.8169	0.5538
1st test		
inv. with discount	0.9148	0.8214
inv. w/o discount	0.8629	0.7659
2nd test		
inv. with discount	0.9501	0.9220
inv. w/o discount	0.9348	0.9061

* Alternative Raw Scores are derived by assuming the input to each question, and checking to see whether the answer is correct for that input.

Table 12

MICROMETER "OURS-THEIRS" SAMPLE: WITHIN SITE RANK CORRELATIONS FOR TRAINEES' PERFORMANCE

101	-0.2
402	0.7
704	0.5
707	0.9
412	0.9
713	0.8
719	0.4
920	0.5

Table 13

INVOICE "OURS-THEIRS" SAMPLE: WITHIN SITE RANK CORRELATIONS FOR TRAINEE PERFORMANCE

Site	Raw score	Alternative score
02	0.0	0.0
03	0.3	0.0
04	0.2	0.4
08	0.0	0.2
09	0.3	0.2
10	-0.5	-0.2
11	0.5	-
14	-0.6	0.4
16	-0.3	0.5
17	0.7	0.1
18	0.2	0.2
19	0.2	0.3